Hamman, W. R., Beaubien, J. M., & Holt, R. W. (1999). Evaluating instructor/evaluator inter-rater reliability from performance database information. Proceedings of the Tenth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University Press.

EVALUATING INSTRUCTOR/EVALUATOR INTER-RATER RELIABILITY FROM PERFORMANCE DATABASE INFORMATION

Captain William R. Hamman
United Air Lines
Denver, CO

J. Matthew Beaubien Robert W. Holt
George Mason University
Fairfax, VA

ABSTRACT

Carriers operating under the FAA's Advanced Qualification Program are required to assess individual and crew performance via Line Oriented Evaluations (LOEs). LOEs take place in a full-motion simulator, and involve a full crew performing a simulated flight from take-off to landing. Evaluating crew performance in the LOE is an arduous task, even for highly-trained professionals. Therefore, techniques are needed for training Instructor/Evaluators (I/Es), and for maintaining I/E calibration indefinitely. This paper describes the major steps involved in the development of Inter-Rater Reliability (IRR) training programs, as well as the usefulness of LOE performance database information for assessing I/E calibration between IRR training sessions.

BACKGROUND

Inter-Rater Reliability (IRR) training programs have been designed to: (1) assist pilot Instructor/Evaluators (I/Es) in determining their strengths and weaknesses as assessors of pilot/crew performance; and (2) reduce various types of rater errors, including personal interpretations of the carrier's Standard Operating Procedure (SOP), memory-based errors, and scale- based errors (see Murphy & Cleveland, 1995 for a detailed description of rater errors and error training).

IRR training is conducted in a group session, during which a cadre of pilot evaluators observes a videotape of crew performance segments, makes independent ratings of each segment, and then discusses the reasons for their differences of opinion. During the course of the training program, subject matter experts (SMEs) provide the I/Es with individual- and group-level feedback regarding their performance in comparison to carrier-specific benchmarks. Such personalized feedback provides I/Es with insight into the way that they typically make performance ratings in the LOE.

At the same time, the discussion that follows assists the cadre of I/Es in reaching some degree of consensus, such that when they return to evaluating crew performance in the simulator, they will be doing so with a common frame of reference (Holt, Johnson, & Goldsmith, 1997; George Mason University, 1996).

The personalized feedback provided to each I/E contains information regarding:

(1) the congruency between each I/E's distribution of judgments and the groups' distribution of judgments
(2) the degree to which each IE's mean performance rating systematically differs from that of the group's overall mean
(3) the degree to which I/Es are able to consistently shift their evaluations upward (when observing better performance) and downward (when observing poor performance) with the group
(4) the degree to which I/Es are able to discriminate between crews of varying performance levels
(5) the absolute level of inter-rater agreement (corrected for chance) on each scale item

When presented independently, such feedback can be misleading. However, feedback regarding all five characteristics -- congruency, systematic differences, consistency, sensitivity, and agreement -- provides the I/Es an in-depth, multi-faceted profile of their strengths and weaknesses as evaluators.

THE COMPONENTS OF DEVELOPING AN IRR TRAINING PROGRAM

The event set (ES) is the primary unit of both CRM assessment and LOE scenario design. An event set consists of a group of related events -- environmental triggers and detailed performance criteria -- that are included in the LOE to assess performance regarding a specific training objective (Federal Aviation Administration, 1990; Hamman, Seamster, Smith, & Lofaro, 1993; Prince, Oser, Salas, & Woodruff, 1993).

Even though an LOE consists of multiple event sets, they are typically linked in such a way as to simulate an uninterrupted flight from start to finish. By segmenting the simulated flight into a small

number of cognitively meaningful "chunks", event sets assist the I/Es in by reducing cognitive workload, and increasing the independence of judgments for each event set.

The first step in the development of IRR training is the identification of performance standards for the primary CRM training objectives, and their integration with the primary technical training objectives. Next, detailed success criteria are developed for evaluating individual and crew performance on each event set. These criteria typically include a set of rules for combining the CRM- and technically-oriented ratings into an overall, crew-level evaluation. An example set of success criteria appears below:

1 Either all observable behaviors for the event set are "Not Observed", OR at least two skills listed for that event set are rated a "1" (Unsatisfactory).

2 Either one observable behavior for the event set is "Not Observed", OR any of the skills for that event set have a "2" rating (Satisfactory).

3 All observable behaviors for the event set are "Fully Observed" or "Partially Observed" (Standard).

4 All observable behaviors are "Fully Observed", AND all skills have a "3" or better rating, with at least one skill rated as "4" (Above Standard)

In addition to basing event set evaluations on the observable behaviors and tasks listed for that event set, general success criteria must also be developed, and considered in the final crew assessment. Typically, general success criteria include the following:

1 The aircraft landed safely.

2 The crew flew within legal limits, or there was appropriate use of emergency authority.

3 The flight remained within guidelines set forth by carrier SOP (or deviations were explained).

4 Appropriate action was taken in a timely manner.

These evaluation criteria provide a number of advantages over other rating techniques. First, the success criteria provide a foundation for the standardization of final judgments. This is critical to the evaluation process. Second, the success criteria are based on objective measurable outcomes. Not only does this allow for the fair evaluation of crew performance, but previous experience suggests that these objective standards are both readily understood and accepted by crews.

Finally, if properly developed, the event sets closely mimic error chains that have been documented in air carrier accidents. This happens because an error during a given event sets does not necessarily have severe consequences *per se*. However, as the crew proceeds to the next event set, prior errors may exacerbate an already complex situation, thereby increasing the chance of failure (Federal Aviation Administration, 1990; Wiener, Kanki, & Helmreich, 1993).

CREATING TRAINING MEDIA FOR THE IRR TRAINING PROGRAM

The second step in the development of IRR training is the creation of videotape examples for the calibration training. If possible, these videotapes should be created with actual line crews flying event set scenarios with no guidance or scripting. Doing so will remove any artificiality caused by the crews' (lack of) acting abilities, and will also provide more realistic examples for evaluators to assess.

For maximum effectiveness, the final videotaped samples of crew performance should be based on performance levels that are rated by SMEs as being marginally safe vs. unsafe. Quite simply, evaluating extreme examples of safe vs. unsafe crews is a rather easy task, and is somewhat unrepresentative of the conditions typically faced by the I/Es. Further, extreme performance examples are likely to be of little use in honing the I/Es' ability to distinguish between crews of similar performance levels.

BUILDING A DATABASE TO ACCEPT LOE SCENARIO OBJECTIVES AND I/E RATINGS OF CREW PERFORMANCE

Crew performance in the LOE is a function of many factors -- including crewmembers' levels of CRM and technical proficiency, the I/Es' skill level, as well as the underlying skill dimensions being assessed in the event set. Given the multitude of factors that can influence crew performance, it is essential that a computerized database be developed to integrate these various sources of influence.

The database must be established to accept LOE scenario objectives, related Terminal Proficiency Objectives (TPOs), primary and secondary CRM categories, and observable crew behaviors for each event set. Further, this database must be linked to the assessment tools used by the I/Es. Doing so will create a complete package for assessment of both I/E calibration and pilot/crew performance, as well as the structural validity of the LOE assessment process.

Data collected during IRR training will be the basis for an Instructor/Evaluator Database (IEDB). The IEDB should also include additional information relevant to the quality of instruction and evaluation. For pilot instructors, this may include instructional qualifications, instructional experience (e.g. classes taught), class evaluations of the instructor, and formal evaluations of the performance of classes taught by the instructor could be included. For evaluators, this may include IRR calibration session results, and comments/ratings by line pilots whom they have evaluated (Beaubien, Holt, & Hamman, 1999).

Creating an integrated repository of information in separate databases will allow carriers the ability to ask difficult questions concerning crew and I/E performance. These questions may include:

1. Why did the percent of pilots failing initial qualification increase this year?
2. How does pilot performance on last year's recurrent LOE point to necessary instructional curriculum changes?
3. What parts of a pilot's training performance during initial qualification predict continuing qualification performance?
4. Which knowledge, skills, or abilities (KSAs) really predict pilot performance?
5. Which training significantly changes these KSAs?
6. To what extent do different types of CRM training experiences predict later line performance?
7. Why are some of the I/Es more effective instructors than others?
8. What additional training would help I/Es with low effectiveness?
9. When have I/Es drifted off calibration benchmarks enough to require remedial IRR calibration training?

Because issues regarding the development of relational databases lie beyond the scope of this document, interested readers are directed to a well-written exposition by Ullman (1982). As noted by Ullman, the construction of a relational database that is based on maximally usable information requires the linking the information from a number of relational data tables.

For example, de-identified PIN numbers may be used to connect pilot background information (prior experience, background, hiring evaluation results, fleet common indoctrination training) to performance at later stages in their tenure with the carrier (qualification results, continuing qualification results, transition training results).

Likewise, this core of pilot background, training and assessment information must be connected to other databases such as the Program Audit database (PADB) and the Instructor/Evaluator database (IEDB). Typically, the PADB is linked to pilot training and evaluation information via systematic content links of curriculum elements and objectives to pilot training (e.g. LOFT) or testing (e.g. LOE, maneuvers validation) events. Similarly, the IEDB is linked to pilot training and evaluation information via I/E identification numbers (Beaubien, Holt, & Hamman, 1999).

APPLICATION OF IRR TO LOE PERFORMANCE DATA

After the LOE has been developed, the tools created, and the cadre of I/Es have been trained, key aspects of IRR calibration can be assessed based on data contained in the LOE performance database. To do this, however, certain operational conditions must exist in practice.

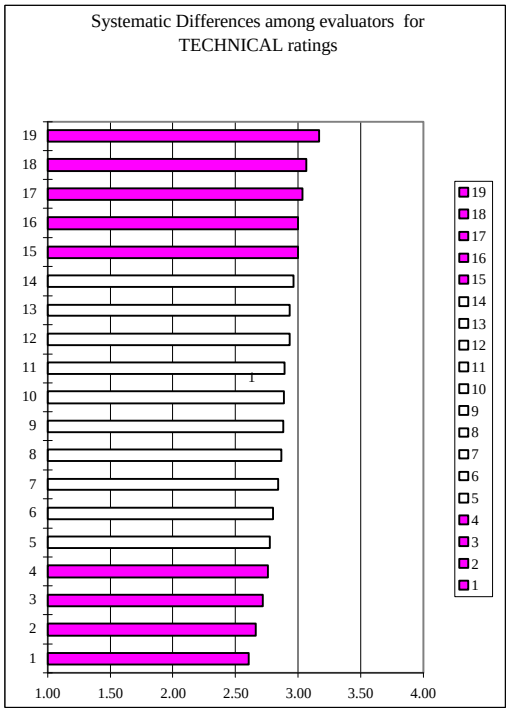
First, I/Es should be matched to crews in a random fashion. Second, each I/E should evaluate a large, representative sample of crews. The combination of these two phenomenon -- random assignment and large, representative samples of pilot/crew performance -- virtually guarantee that when statistically and practically significant differences are observed across I/Es, they are a function of systematic rater characteristics, rather than idiosyncratic characteristics of the sample of ratees.

If these conditions exist in practice, three aspects of IRR training (systematic differences, congruency, and consistency) can be assessed. An example of these three IRR benchmarks are as follows:

SYSTEMATIC DIFFERENCES

Systematic differences in mean ratings among the group of raters is indicated by an Analysis of Variance (ANOVA) using the rater as a factor. The systematic difference of each rater from the group average is heuristically estimated by a t-test of the rater's average vs. the group's average.

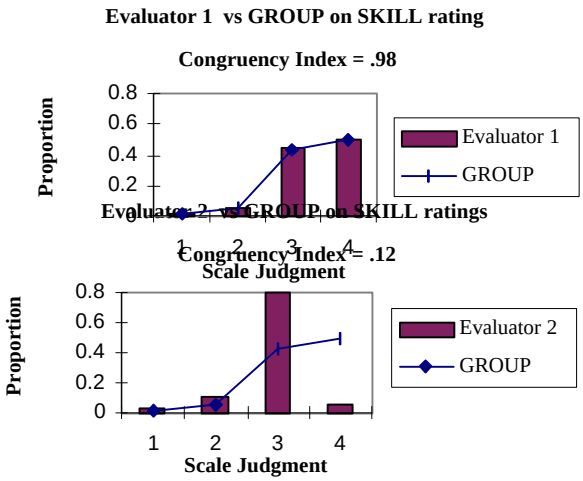
The overall amount of variance in I/Es' ratings may be due to a number of factors, including evaluator systematic differences (undesirable), event set systematic differences (desirable), evaluator by event set interactions (undesirable). Typically, a pie chart is used to illustrate the percent of variance explained by each source, while a bar chart is used to indicate each evaluator's mean evaluation in comparison to the group average. In general, I/Es with statistically significant high or low mean ratings are "red-flagged" to catch management's attention. An example appears below.



CONGRUENCY

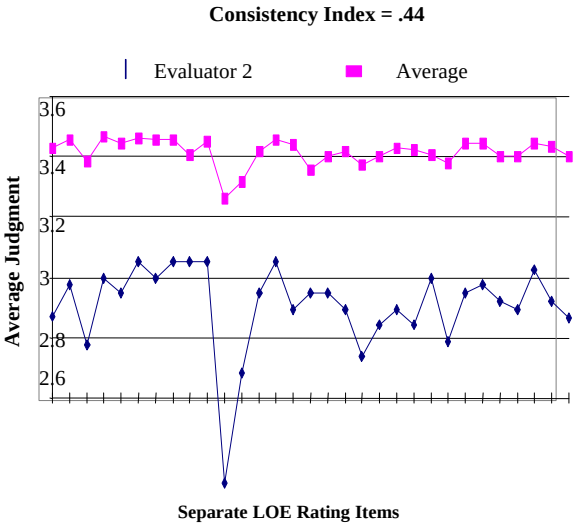
Measures of congruency assess the shape of the distribution of ratings of each I/E (individually) compared to the group's distribution. They complement ratings of systematic differences by suggesting how each I/E's mean rating came about. The Congruency Index (CI) is calculated by comparing group &

individual probabilities: CI = 1 - Σ |(P_i - P_g)|, where P_i equals the relative proportion of evaluator *i*'s ratings occurring at that scale point, and P_g equals the relative proportion of group's ratings occurring at that scale point (George Mason University, 1996). The results of the congruency index are presented in graphic form, with ratings that range from 0.0 (no congruency) to 1.0 (perfect congruency). Two examples appear below. The first represents a high level of congruency; the second represents a low level of congruency.



CONSISTENCY

Rater consistency is indexed by how much raters' evaluations intercorrelate. Conceptually, the inter-rater correlation is the extent to which raters consistently shift upward (when observing better performance) and downward (when observing poorer performance) with the group. More specifically, each I/E's rating profile (across items) is compared to the overall group profile, and the consistency index is calculated using the Pearson product-moment correlation (r). The consistency graph below shows the individual and group judgment profiles across items on a given LOE. The consistency index is based on the shape of the two distributions and ignores mean differences in judgments. Therefore, the two distributions below show moderately consistent patterns of judgments.



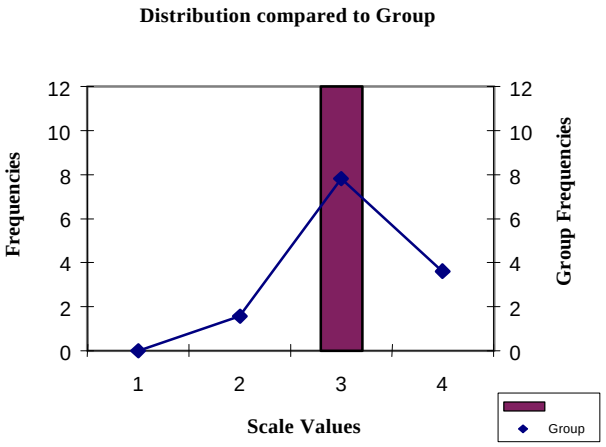
REAL-LIFE EXAMPLES

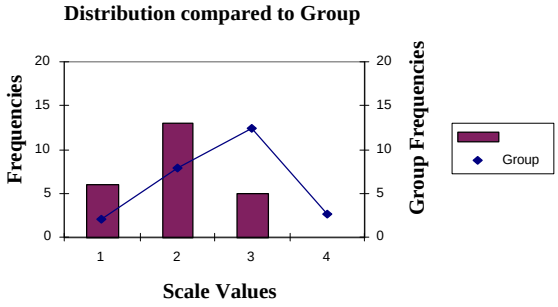
The monitoring and use of this statistically-driven information is critical for maintaining the overall quality of the assessment process. Precise, high-quality answers require sensitive, reliable, and valid data. Obtaining this type of data is facilitated by well-developed research design, content- and construct-valid measures, and the training of evaluators via IRR training.

The IRR training classes and ongoing monitoring have identified several common "evaluation profiles". These include the "Midline Evaluator", the "Easy Evaluator", the "Hard Evaluator", and the "Good Evaluator". In the following sections, characteristics of these raters will be discussed in more detail. The examples are based upon real data, but have been de-identified to ensure anonymity.

The Midline Evaluator

The mid-line rater is very common among groups of evaluators who do evaluate crew performance full-time, such as domicile personnel. The fundamental reason for midline assessment is the raters' feelings of unease with the assessment criteria, and their aversion to making mistakes. As a result, a substantial portion of midline evaluators' ratings are "3" (Standard performance). Once identified, however, these individuals typically respond well to training. An example appears below.



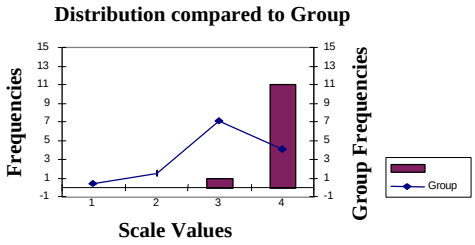


The Easy Evaluator ("Santa Claus")

This evaluation profile is typically found among groups of evaluators who do not want to put forth the effort to understand the performance standards, or among individuals who have been evaluators for an extremely long period of time. Their continued exposure to crews has caused them to shift their assessment to a comparison with other crews rather than a comparison to the carrier-specific performance standards. As a result, a substantial portion of these evaluators' ratings are "4" (Above Standard).

This is the most dangerous group of evaluators, because crews exposed to this evaluator may be allowed to fly, even though they are (actually) of substandard performance. Rarely is negative feedback provided

concerning this type of evaluator. Therefore, fleet personnel and/or quality assurance may not be aware of these individuals. An example appears below.



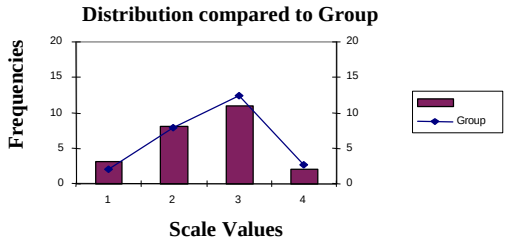
The Hard Evaluator ("Ax Man")

The Ax-man evaluator is known by everyone. This group of evaluators usually has the problem of being strongly biased by one event during the check ride, thereby causing their assessments to be extremely harsh for the remainder of the evaluation. . As a result, a substantial portion of these evaluators' ratings are "1" (Unsatisfactory) or "2" (Satisfactory). Such individuals, usually have a hard time being objective during the evaluation.

If their biasing is extreme, this group will typically respond poorly to the IRR training, and will continue to perform poor evaluations for the remainder of their tenure as an evaluator. There is a tendency for new evaluators to rate harsher in their assessments as a result of their application of existing standards. This group should not be confused with the true ax-man as this group will respond well to training and become excellent evaluators. An example appears below.

The Good Evaluator

The good evaluator is someone who has a reasonable understanding of the assessment standards, can apply these standards across crews in a equitable fashion, and can maintain objectivity even if the crew/pilot makes errors during the assessment. In other words, mistakes do not bias these evaluators' ability to assess the pilot/crew on other objectives of the assessment. These groups of evaluators not only make good evaluations, but they are also good facilitators of the IRR training. They can train by example, and model the characteristics of an excellent evaluation for others.



DISCUSSION

In recent years, inter-rater reliability (IRR) training programs have been developed to assist I/Es in understanding their strength and weaknesses as evaluators (George Mason University, 1996). This tool has been extremely valuable in the initial training of I/Es, and is now becoming a useful technique for monitoring the ongoing calibration of evaluators so that problems can be identified before they become catastrophic.

The IRR process identifies confusion about operating standards, interpretation problems with assessment forms, and degradation of assessment performance by individual evaluators. Additionally, the IRR process identifies "profiles" of evaluators that may be used in the future to select evaluators.

The different types of evaluation profiles previously discussed are in some degree inherent to the individual. The IRR training can improve and shift ratings to some degree, but if an individual is extreme in their assessment profiles, the IRR training may have little impact. Because of this, it would be beneficial to create a selection type of IRR process which will measure the inherent assessment profile of an individual. An individual with a extreme rating either lenient or harsh should therefore be excluded from further consideration as a potential evaluator.

REFERENCES

Beaubien, J. M., Holt, R. W., & Hamman, W. R. (1999). Evaluating LOE quality from performance database information. Proceedings of the Tenth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University Press.

Federal Aviation Administration. (1990). Line operational simulations: Line oriented flight training, special purpose operational training, line oriented evaluation. Advisory Circular 120-35B. Washington, DC: Author.

George Mason University. (1996, November). Improving crew assessments. Training materials to accompany an FAA sponsored workshop on evaluator calibration. FAA Grant Team. George Mason University, Fairfax, VA: Author.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1997). Application of psychometrics to the calibration of air carrier evaluators. Unpublished manuscript. George Mason University, Fairfax, VA.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance appraisal: Social, cognitive, and goal-based perspectives. Newbury Park, CA: Sage.

Prince, C., Oser, R., Salas, E., & Woodruff, W. (1993). Increasing hits and reducing misses in CRM/LOS scenarios: Guidelines for simulator development. International Journal of Aviation Psychology, 3(1), 69-82.

Wiener, E. L., Kanki, B. G., & Helmreich, R. L. (1993). Cockpit resource management. San Diego, CA: Academic Press.